(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110619042 B
(45)授权公告日 2022.08.16

(21)申请号 201911066052.8
(22)申请日 2019.11.04
(65)同一申请的已公布的文献号
申请公布号 CN 110619042 A
(43)申请公布日 2019.12.27
(66)本国优先权数据
201910187119.7 2019.03.13 CN
(73)专利权人 北京航空航天大学
地址 100191 北京市海淀区学院路37号
(72)发明人 吴文峻 吴沂楠
(74)专利代理机构 北京科迪生专利代理有限责任公司 11251
专利代理师 安丽 邓治平
(51)Int.Cl.
G06F 16/332(2019.01)
G06N 3/04(2006.01)
(56)对比文件
CN 105760417 A,2016.07.13
CN 109062958 A,2018.12.21
US 2009186330 A1,2009.07.23
CN 108021703 A,2018.05.11
EP 1686513 A2,2006.08.02
CN 101000600 A,2017.07.18
CN 108288468 A,2018.07.17
高翔.基于神经网络的问答系统关键技术研究.《中国优秀博硕士学位论文全文数据库(硕士)信息科技辑》.2018,(第10期),I138-946.

Kavya Kishore等.Document Summarization in Malayalam with sentence framing.《2016 International Conference on Information Science (ICIS)》.2017,第194-200页.

吴文峻.面向智慧教育的学习大数据分析技术.《电化教育研究》.2017,第38卷(第06期),第88-94页.

曾帅等.面向知识自动化的自动问答研究进展.《自动化学报》.2017,第43卷(第09期),第1491-1508页.
审查员 雷婉婧

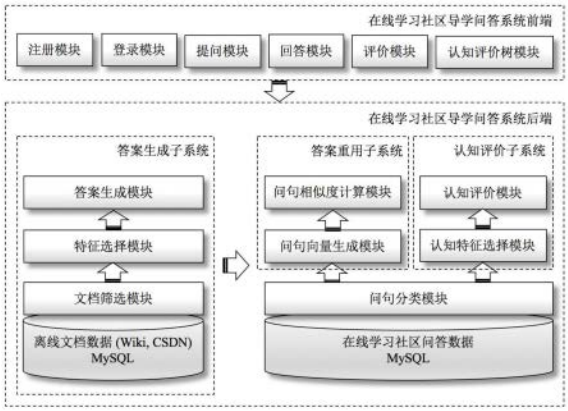
权利要求书2页 说明书5页 附图3页

(54)发明名称

一种基于神经网络的导学问答系统及方法

(57)摘要

本发明设计了一种基于神经网络的导学问答系统,包括前端和后端,前端包括注册模块、登录模块、提问模块、回答模块、评价模块和认知评价树模块;所述后端包括三个子系统,分别为认知评价子系统、答案重用子系统和答案生成子系统;本发明通过对问句的信息提取,检索社区中已存在的问题的答案;如果不存在相似的问题,则使用训练好的模型在离线文档中生成回答,并在返回的答案中引导学习者扩展阅读。



CN 110619042 B

1.一种基于神经网络的导学问答系统,其特征在于:包括前端和后端,前端包括注册模块、登录模块、提问模块、回答模块、评价模块和认知评价树模块;所述后端包括三个子系统,分别为认知评价子系统、答案重用子系统和答案生成子系统;

注册模块,提供学习者注册的功能,学习者必须注册成功后才能登录;

登陆模块,提供学习者登录的功能,学习者必须登录成功后才能使用提问模块、回答模块和评价模块的功能;

提问模块,提供学习者提问的功能,学习者在提问界面撰写问题标题、问题描述和问题分类;

回答模块,提供学习者回答其他学习者的问题的功能,学习者在回答界面撰写回答内容;

评价模块,提供学习者评价其他学习者的回答功能,学习者点赞、反对标记为标准答案或者在评价界面撰写评价内容中;

评价树模块,提供教师浏览课程中发帖认知程度的功能;

认知评价子系统,用于对学习者发帖的认知行为进行评价,最后为每个概念和知识点展示一棵认知评价树,为教师干预在线学习社区中学习者的对话提供依据;认知评价子系统包括认知特征选择模块和认知评价模块;认知特征选择模块根据学生的发帖内容,对认知进行分类,并得到学生认知状态的特征向量;认知评价模块根据学生认知状态,得到认知评价树;

答案重用子系统,将在线学习社区中已存在的含有高质量答案的问题进行重复利用,减少答案的冗余并快速相应学习者的提问;答案重用子系统包括:问句向量生成模块和问句相似度计算模块,问句向量生成模块根据问题的文本内容,得到问题的向量表示;问句相似度计算模块根据问题的向量表示,得到问题间相似度的接近程度;

答案重用子系统的流程描述为:

(1)对学习者提出的问题进行文本上的预处理;

(2)对于预处理完成的问题通过布尔规则进行分类,根据分类在问题集中找到相同分类的、包含有高质量答案的问题的目标问题集合;

(3)对预处理完成的问题中的词语进行语义扩展,以丰富问句的关键词数;

(4)对目标问题集合中的问题和扩展后的问句通过TF-IDF进行关键词的加权计算得到特征向量;

(5)将目标问题集合中的问题向量和学习者的问题向量进行相似度计算;将相似度最高的问题下点赞数最多或者被标记为标准答案的回答返回给学习者,同时返回改问题的URL链接,引导学习者阅读该相似的问题;

答案生成子系统,在无法找到相似问题时根据离线的文档数据对学习者的提问生成回答;答案生成子系统包括文档筛选模块、特征选择模块和答案选择模块;文档筛选模块根据文档间的相似程度,得到相近的文档集合;特征选择模块根据帖子的文本内容,得到帖子的数字向量表示;答案选择模块根据神经网络的输出,得到可能作为答案的片段;

答案生成子系统的流程描述为:

(1)对学习者提出的问题进行预处理,此时需要保留部分标点符号作为特征;

(2)对预处理后的问题进行n-gram扩展并将每一个词语哈希到二进制中的一位中;

(3)将二进制编码在文档集合中进行检索,筛选出最相关的文档集合;

(4)对问题和文档分别提取特征,得到输入神经网络的向量;

(5)将问题向量和文档向量分别输入两个多层双向的RNN网络,得到问题和文档的隐含层特征向量;

(6)将问题和文档的隐含层特征向量输入双线性分类器以预测答案开始和结束的位置,将生成的答案返回给用户,并附上文档来源的链接,以引导学习者进入链接扩展阅读;

前端界面中的模块和后端系统中的各个模块通过HTTP协议进行交互。

2.一种实现基于权利要求1所述系统的基于神经网络的导学问答方法,其特征在于,首先在特征上增加了短语的属性特征,这些特征的引入使得系统不仅能从语义上表示短语,更能从词性、实体信息上为短语打上标签,然后在注意力机制方面,加入了对问题的自对齐注意力特征,这使得模型不仅能根据问题对文档中的词语加入权重,也能根据问题中词语的重要程度在训练过程中更新权值;

具体实现步骤为:

(1)对于学习者的提问,首选使用布尔表达式进行概念和知识点的分类,对于分好类的问题,并行地进入答案重用子系统和认知评价子系统中;

(2)在答案重用子系统中,根据问题的知识点分类筛选出同类并且具有回答并且回答被其他学习者赞同或标记为答案的问题,并和学习者的提问进行相似度计算,筛选出最相似的问题,并返回该问题下点赞数最多或者被标记为标准答案的回答;

(3)如果不存在相似问题的答案,则进入答案生成子系统中,根据离线文档生成问题的答案,如果该答案评分大于给定的阈值,则认为答案质量高,并返回答案;

(4)如果问题没有学习者回答,或者回答没有被赞同或标记为答案,则放弃自动回答,等待其他学习者对提问进行回答;

(5)通过认知评价子系统生成的认知评价树,教师对在线学习社区对话进行干预,从而引导学习者学习;通过答案重用子系统和答案生成子系统返回的答案,在答案中将加入答案来源以引导学习者扩展阅读。

一种基于神经网络的导学问答系统及方法

技术领域

[0001] 本发明涉及一种基于神经网络的导学问答系统及方法,属于神经网络和自动问答领域。

背景技术

[0002] 学习科学和计算机辅助学习领域的研究提供了证据表明,学习者的对话是其建构知识网络的重要影响因素。在线下教学中,课堂问答是传统线下教学模式中非常重要的组成部分,因此在线学习社区的在线问答技术还原了线下教学的场景,为学习者和老师提供了一个与学习伙伴互动的在线环境,促进了社区中老师与学习者及学习者与学习者之间的知识流动和经验共享。然而,在线问答缺乏实时性,例如由斯坦福大学创办的MOOC平台“Coursera”的问答平台中,学习者提出的问题得到回答的中值时间达22分钟。这使得学习者不能及时解决疑惑,使得教师不能及时了解学习者对知识点的掌握情况,无法达到线下教学“面对面”及时答疑的模式。

[0003] 针对在线问答实时性的弊端,自动问答技术的引入允许学习者通过自然语言进行叙述,系统结合自然语言处理技术“理解”学习者的问题,并检索得到相应的答案实时返回给学习者。但是由于在线学习社区学习者的不断增加,学习者在社区中提出的问题不能及时、准确的得到其他用户的回答。同时,由于不同学习者对于问题的描述不尽相同,使得相同的并且已经存在答案的问题不能及时反馈给学习者。然而,仅仅将问题的答案反馈给学习者只能使其被动的接受答案,而不能使其主动地扩展阅读,也不能反映学习者参与问题讨论的认知变化。因此,需要一种能够既能及时准确解决学习者的疑惑、也能引导学习者扩展阅读,最后将学习者认知的变化及时反映的问答系统,即导学问答系统,来解决上述问题。

发明内容

[0004] 本发明的技术解决问题是:克服了在线问答实时性和准确性的弊端,提供一种基于神经网络的导学问答系统及方法,具有能够既能及时准确解决学习者的疑惑、也能引导学习者扩展阅读,最后将学习者认知的变化及时反映的优点。

[0005] 本发明技术解决方案:

[0006] 一种基于神经网络的导学问答系统,包括前端和后端,前端包括注册模块、登录模块、提问模块、回答模块、评价模块和认知评价树模块;所述后端包括三个子系统,分别为认知评价子系统、答案重用子系统和答案生成子系统;

[0007] 注册模块,提供学习者注册的功能,学习者必须注册成功后才能登录;

[0008] 登陆模块,提供学习者登录的功能,学习者必须登录成功后才能使用提问模块、回答模块和评价模块的功能;

[0009] 提问模块,提供学习者提问的功能,学习者在提问界面撰写问题标题、问题描述和问题分类;

[0010] 回答模块，提供学习者回答其他学习者的问题的功能，学习者在回答界面撰写回答内容；

[0011] 评价模块，提供学习者评价其他学习者的回答功能，学习者点赞、反对标记为标准答案或者在评价界面撰写评价内容中；

[0012] 评价树模块，提供教师浏览课程中发帖认知程度的功能；

[0013] 认知评价子系统，用于对学习者发帖的认知行为进行评价，最后为每个概念和知识点展示一棵认知评价树，为教师干预在线学习社区中学习者的对话提供依据；认知评价子系统包括认知特征选择模块和认知评价模块；认知特征选择模块根据学生的发帖内容，得到学生认知状态的特征向量；认知评价模块根据学生认知状态，得到认知评价树；

[0014] 答案重用子系统，将在线学习社区中已存在的含有高质量答案的问题进行重复利用，减少答案的冗余并快速相应学习者的提问；答案重用子系统包括：问句向量生成模块和问句相似度计算模块，问句向量生成模块根据问题的文本内容，得到问题的向量表示；问句相似度计算模块根据问题的向量表示，得到问题间相似度的接近程度；

[0015] 答案生成子系统，在无法找到相似问题时根据离线的文档数据对学习者的提问生成回答；答案生成子系统包括文档筛选模块、特征选择模块和答案选择模块；文档筛选模块根据文档间的相似程度，得到相近的文档集合；特征选择模块根据帖子的文本内容，得到帖子的数字向量表示；答案选择模块根据神经网络的输出，得到可能作为答案的片段；

[0016] 前端界面中的模块和后端系统中的各个模块通过HTTP协议进行交互。

[0017] 本发明的一种基于神经网络的导学问答方法，步骤为：

[0018] (1)对于学习者的提问，首选使用布尔表达式进行概念和知识点的分类，对于分好类的问题，并行地进入答案重用子系统和认知评价子系统中；

[0019] (2)在答案重用子系统中，根据问题的知识点分类筛选出同类并且具有回答并且回答被其他学习者赞同或标记为答案的问题，并和学习者的提问进行相似度计算，筛选出最相似的问题，并返回该问题下点赞数最多或者被标记为标准答案的回答；

[0020] (3)如果不存在相似问题的答案，则进入答案生成子系统中，根据离线文档生成问题的答案，如果该答案评分大于给定的阈值，则认为答案质量高，并返回答案；

[0021] (4)如果问题没有学习者回答，或者回答没有被赞同或标记为答案，则放弃自动回答，等待其他学习者对提问进行回答；

[0022] (5)通过认知评价子系统生成的认知评价树，教师对在线学习社区对话进行干预，从而引导学习者学习；通过答案重用子系统和答案生成子系统返回的答案，在答案中将加入答案来源以引导学习者扩展阅读。

[0023] 本发明与现有技术相比的优点在于：

[0024] (1)本发明细分为答案重用和答案生成两大模块，通过改进TF-IDF检索算法查找最相似的问题集合，同时通过神经网络方法在离线文档中寻找最可能的答案，结合两者的结果为学习者的提问提供及时准确的回答；

[0025] (2)本发明的认知分类框架能将不同认知细分为具体的对话行为，并结合社区对话内容的特点对每一个认知行为进行准确定义，使得分类模型能够对在线学习社区中的对话进行准确的认知分类；

[0026] (3)本发明使用神经网络模型，对学习者的提问进行语义分析，根据语义在离线文

档中检索相关的内容，然后提取文档和问题的词向量和文本特征，最后通过多层双向循环神经网络得到深层次的隐含层向量特征，最后通过双线性分类器找到答案在文档中出现的位置并返回给学习者。我们通过对答案生成建模的实验验证了模型对比于两个基准模型和两个改进模型的效果提升。

附图说明

[0027] 图1为本发明的系统的组成框图；

[0028] 图2为本发明认知评价子系统实现过程图；

[0029] 图3为本发明答案重用子系统实现过程图；

[0030] 图4为本发明答案生成子系统实现过程图。

具体实施方式

[0031] 下面结合附图及实施过程进行详细说明。

[0032] 如图1所示，

[0033] 本发明系统的模型和框架的编写是在MacOS 10.12.6平台下的PyCharm IDE使用Python语言和Scala语言开发完成。深度学习的模型等训练是在Ubuntu 16.04操作系统下的GTX 1080+Cuda 8.0环境下完成。在神经网络模型实现上，使用了目前较流行的PyTorch深度学习框架。

[0034] 采集学习者的提问，使用布尔表达式进行概念和知识点的分类，对于分好类的问题，并行地进入答案重用子系统和认知评价子系统中；在答案重用子系统中，根据问题的知识点分类筛选出同类并且具有高质量回答的问题，并和学习者的提问进行相似度计算，筛选出最相似的问题，并返回该问题下点赞数最多或者被标记为标准答案的回答；如果不存在相似问题的答案，则进入答案生成子系统中，根据离线文档生成问题的答案，如果该答案评分大于给定的阈值，则认为答案质量高，并返回答案；如果答案质量低，则放弃自动回答，等待其他学习者对提问进行回答；最后，通过认知评价子系统生成的认知评价树，教师可以对社区对话进行干预，从而引导学习者学习；通过答案重用子系统和答案生成子系统返回的答案，在答案中将加入答案来源以引导学习者扩展阅读。

[0035] 后端主要分为三个子系统，分别为认知评价子系统、答案重用子系统和答案生成子系统，各个子系统的功能为：

[0036] (1)认知评价子系统

[0037] 认知评价子系统主要包括认知特征选择模块和认知评价模块，主要用于对学习者发帖的认知行为进行评价，最后为每个概念和知识点展示一棵认知评价树，为教师干预在线学习社区中学习者的对话提供依据；

[0038] (2)答案重用子系统

[0039] 答案重用子系统主要包括问句向量生成模块和问句相似度计算模块，目的是将在线学习社区中已存在的含有高质量答案的问题进行重复利用，减少答案的冗余并快速相应学习者的提问；

[0040] (3)答案生成子系统

[0041] 答案生成子系统主要包括文档筛选模块、特征选择模块和答案选择模块，目的是

在答案重用子系统无法找到相似问题时根据离线的文档数据对学习者的提问生成回答。

[0042] 上述前端界面中的模块和后端系统中的模块通过HTTP协议进行交互,具体交互请求将在后面对每个子系统的设计与实现中阐述。

[0043] 如图2所示,答案重用的流程可以描述为:

[0044] (1)对学习者提出的问题进行文本上的预处理,如取出标点符号、停用词等;

[0045] (2)对于预处理完成的问题通过布尔规则进行分类,根据分类在问题集中找到相同分类的、包含有高质量答案的问题的目标问题集合;

[0046] (3)对预处理完成的问题中的词语进行语义扩展,以丰富问句的关键词数;

[0047] (4)对目标问题集合中的问题和扩展后的问句通过TF-IDF进行关键词的加权计算得到特征向量;

[0048] (5)将目标问题集合中的问题向量和学习者的问题向量进行相似度计算;将相似度最高的问题下点赞数最多或者被标记为标准答案的回答返回给学习者,同时返回改问题的URL链接,引导学习者阅读该相似的问题。

[0049] 如图3所示,答案生成的流程可以描述为:

[0050] (1)对学习者提出的问题进行预处理,此时需要保留部分标点符号作为特征;

[0051] (2)对预处理后的问题进行n-gram扩展并将每一个词语哈希到二进制中的一位中;

[0052] (3)将二进制编码在文档集合中进行检索,筛选出最相关的文档集合;

[0053] (4)对问题和文档分别提取特征,得到输入神经网络的向量;

[0054] (5)将问题向量和文档向量分别输入两个多层双向的RNN网络,得到问题和文档的隐含层特征向量;

[0055] (6)将问题和文档的隐含层特征向量输入双线性分类器以预测答案开始和结束的位置,将生成的答案返回给用户,并附上文档来源的链接,以引导学习者进入链接扩展阅读。

[0056] 如图4所示,本发明将学习者提出问题到系统返回答案分为以下几个步骤:

[0057] (1)导学问答系统对于学习者的提问,首选使用布尔表达式进行概念和知识点的分类,对于分好类的问题,并行地进入答案重用子系统和认知评价子系统中;

[0058] (2)在答案重用子系统中,根据问题的知识点分类筛选出同类并且具有高质量回答的问题,并和学习者的提问进行相似度计算,筛选出最相似的问题,并返回该问题下点赞数最多或者被标记为标准答案的回答;

[0059] (3)如果不存在相似问题的答案,则进入答案生成子系统中,根据离线文档生成问题的答案,如果该答案评分大于给定的阈值,则认为答案质量高,并返回答案;

[0060] (4)如果答案质量低,则放弃自动回答,等待其他学习者对提问进行回答;通过认知评价子系统生成的认知评价树,教师可以对社区对话进行干预,从而引导学习者学习;通过答案重用子系统和答案生成子系统返回的答案,在答案中将加入答案来源以引导学习者扩展阅读。

[0061] 表1展示了本发明使用默认的模型参数和特征构建的答案生成系统以及比较系统在验证集上的预测结果。

[0062] 表1答案生成模型预测结果对比

[0063]

系统	F1（%）
BiDAF 问答系统	70.04
A3Net 问答系统	72.51
导学问答系统	73.31

[0064] 从表中可以看出，其它系统相较于基准模型，F1均有不同程度的提升，其原因在于这些模型改进了模型结构并且引入了新的文本语义表达特征。F1值是精确率和召回率的调和均值，即F1＝2PR/P+R，P为精确率，是预测正确的正样本数和所有正样本数的比值，在该实验中是指系统回答正确的问题数占系统提供了回答的问题数的比值；R为召回率，是预测正确的正样本数和所有判断为正样本数的比值，在该实验中是指系统回答正确的问题数占所有问题数的比值。F1值是评价系统预测效果的重要指标之一。

[0065] 导学问答系统相比于BiDAF系统和A3Net系统，首先在特征上增加了短语的属性特征，这些特征的引入使得系统不仅能从语义上表示短语，更能从词性、实体信息上为短语打上标签。然后在注意力机制方面，加入了对问题的自对齐注意力特征，这使得模型不仅能根据问题对文档中的词语加入权重，也能根据问题中词语的重要程度在训练过程中更新权值。最后，在对输出的隐含层向量映射上，使用了两个线性分类器分别预测答案的起始和结束位置，而不是使用一个softmax分类器，这使得预测之间不受干扰，提高了预测的准确性。综上，导学问答系统相比于BiDAF系统和A3Net系统在F1值上有所提升，这表明系统对学习者的提问进行回答的准确度提高了。

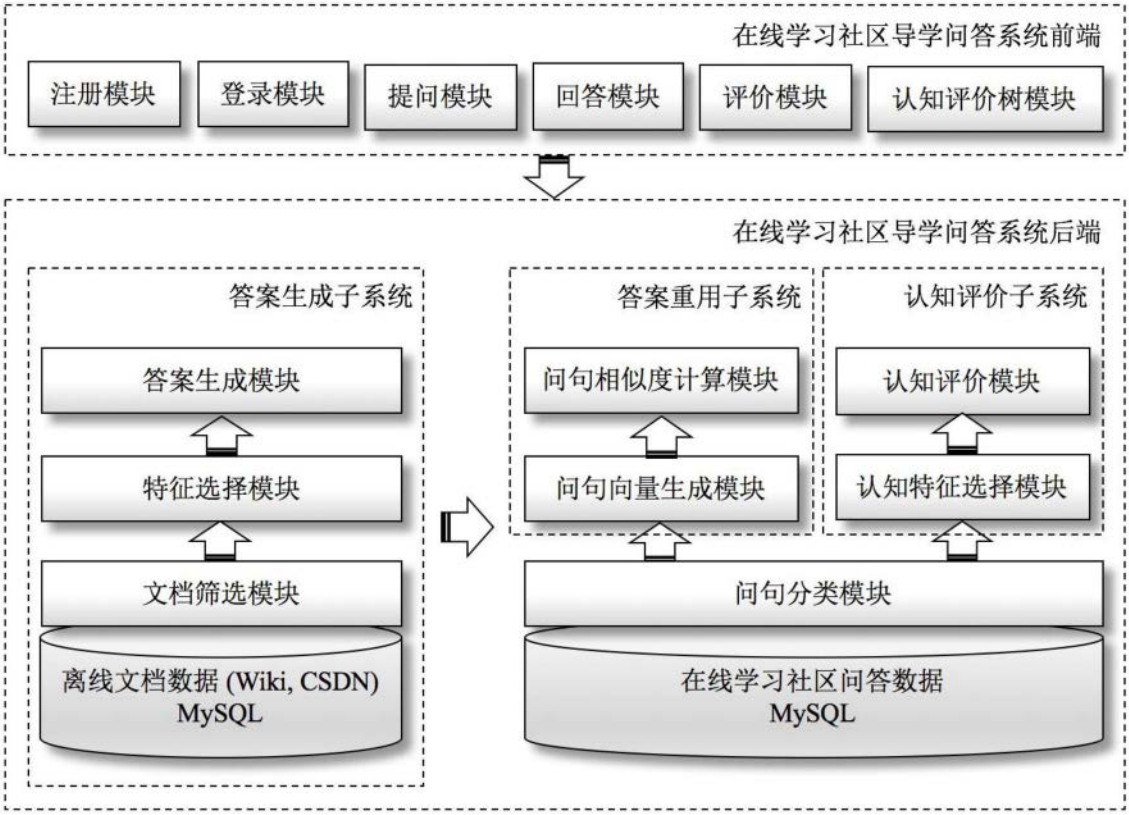

在线学习社区导学问答系统前端
注册模块
登录模块
提问模块
回答模块
评价模块
认知评价树模块
在线学习社区导学问答系统后端
答案生成子系统
答案生成模块
特征选择模块
文档筛选模块
离线文档数据 (Wiki, CSDN)
MySQL
答案重用子系统
问句相似度计算模块
问句向量生成模块
问句分类模块
在线学习社区问答数据
MySQL
认知评价子系统
认知评价模块
认知特征选择模块

图1

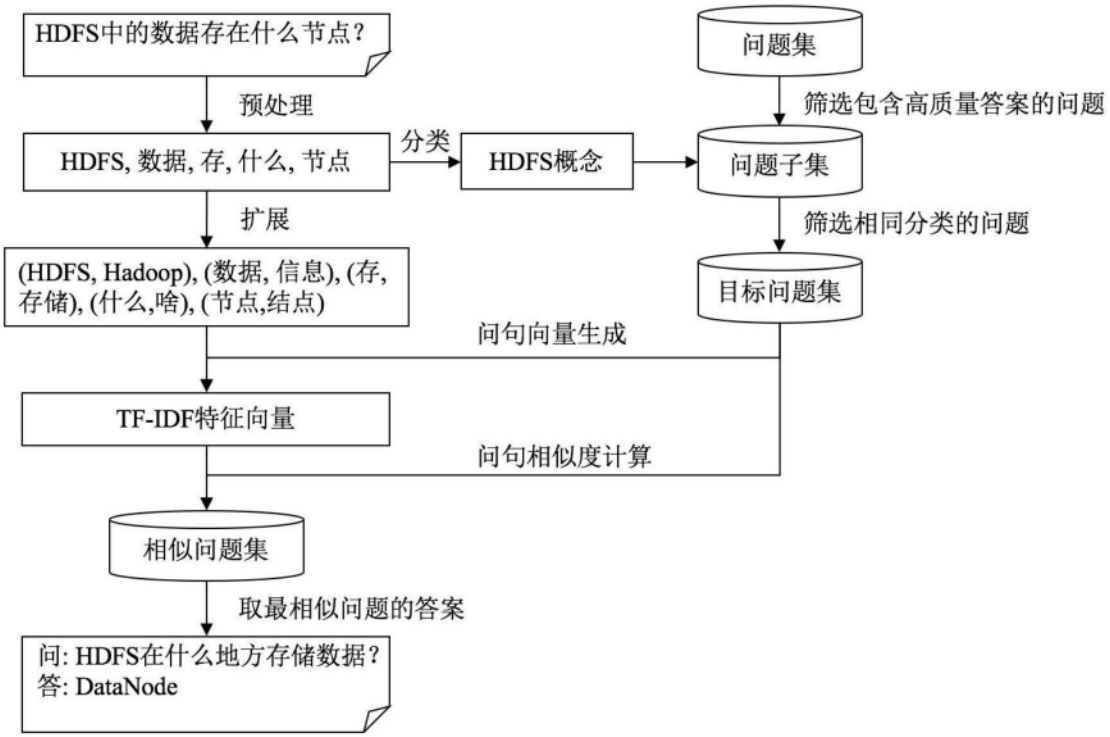

HDFS中的数据存在什么节点?
预处理
HDFS, 数据, 存, 什么, 节点
分类
HDFS概念
扩展
(HDFS, Hadoop), (数据, 信息), (存, 存储), (什么,啥), (节点,结点)
TF-IDF特征向量
相似问题集
取最相似问题的答案
问: HDFS在什么地方存储数据?
答: DataNode
问题集
筛选包含高质量答案的问题
问题子集
筛选相同分类的问题
目标问题集
问句向量生成
问句相似度计算

图2

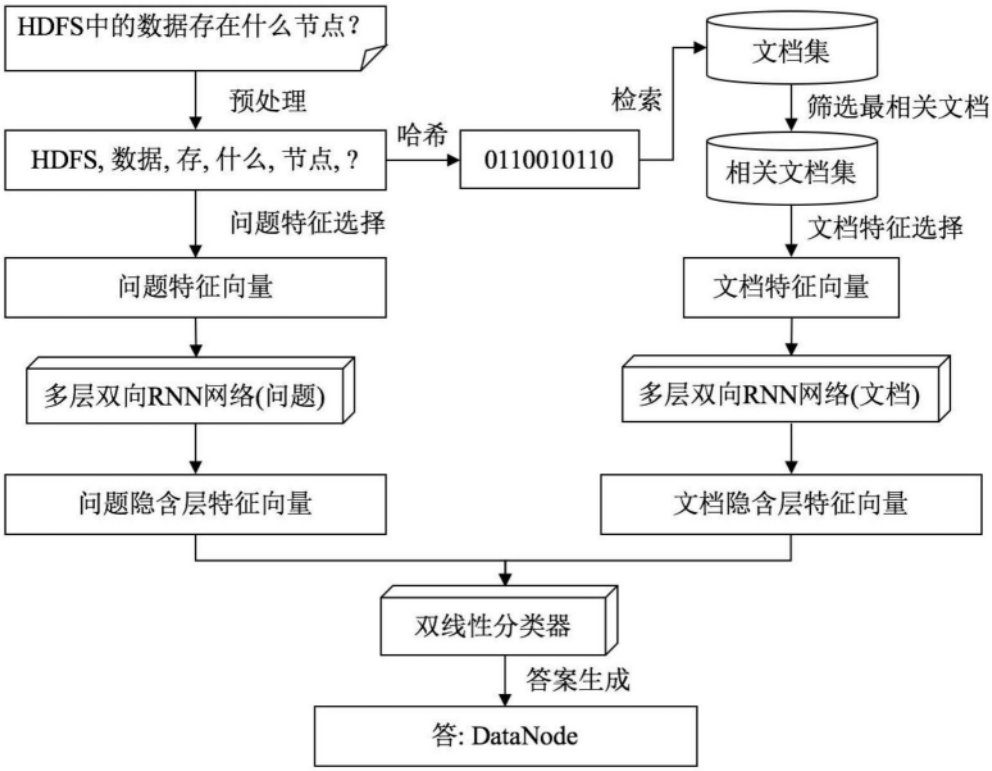

HDFS中的数据存在什么节点?
预处理
HDFS, 数据, 存, 什么, 节点, ?
哈希
0110010110
检索
文档集
筛选最相关文档
相关文档集
问题特征选择
文档特征选择
问题特征向量
文档特征向量
多层双向RNN网络(问题)
多层双向RNN网络(文档)
问题隐含层特征向量
文档隐含层特征向量
双线性分类器
答案生成
答: DataNode

图3

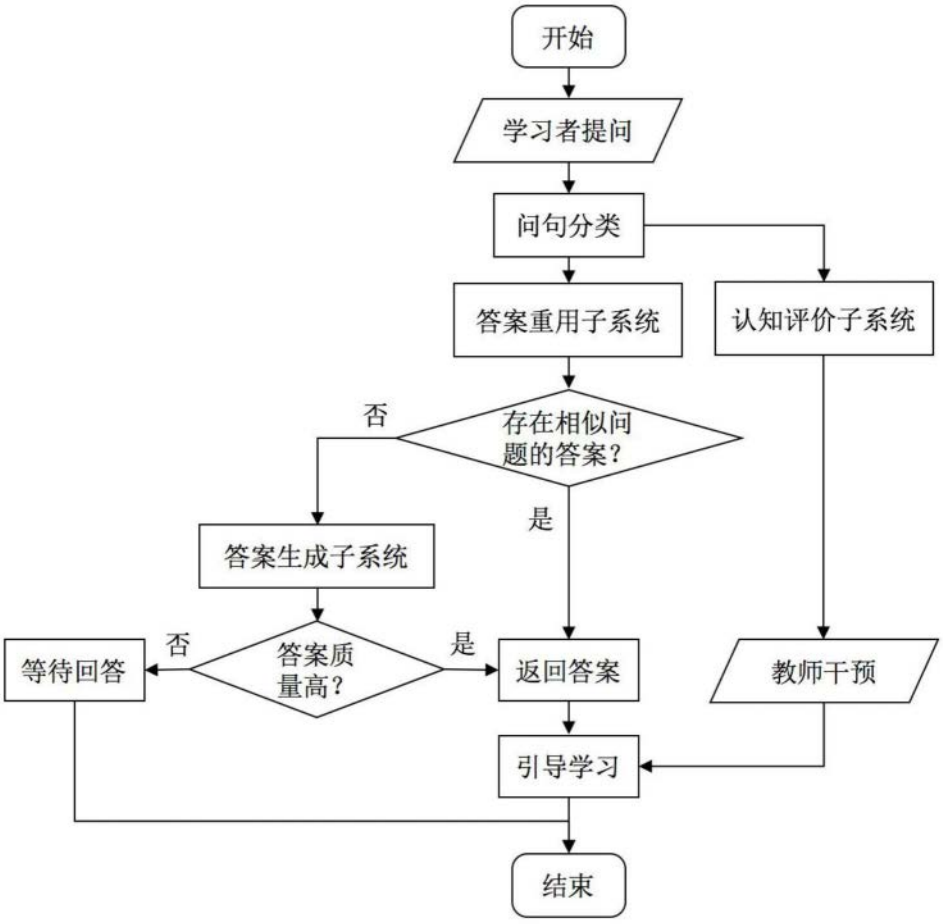

开始
学习者提问
问句分类
答案重用子系统
认知评价子系统
否
存在相似问题的答案?
是
答案生成子系统
否
答案质量高?
是
等待回答
返回答案
教师干预
引导学习
结束

图4